SCB STATISTISKA CENTRALBYRÅN

86-03-20

PROMEMORIOR FRÅN P/STM

NR 22

QUALITY CONTROL OF CODING OPERATIONS AT STATISTICS SWEDEN

AV LARS LYBERG

INLEDNING

TILL

Promemorior från P/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1978-1986. – Nr 1-24.

Efterföljare:

Promemorior från U/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1986. – Nr 25-28.

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

Promemorior från P/STM 1986:22. Quality control of coding operations at Statistics Sweden / Lars Lyberg.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-PM-PSTM-1986-22

SCB STATISTISKA CENTRALBYRÅN

86-03-20

PROMEMORIOR FRÅN P/STM

NR 22

QUALITY CONTROL OF CODING OPERATIONS AT STATISTICS SWEDEN

AV LARS LYBERG

1986-03-20

Quality Control of Coding Operations at Statistics Sweden

Lars Lyberg, Statistics Sweden

Abstract

Coding is a major operation in such statistical studies as censuses of population, censuses of business, and labor force surveys. As with most other survey operations, coding is highly susceptible to errors. Furthermore, coding is difficult to control, because it requires a lot of judgement on the part of the coder. Even experienced coders display a great deal of variation in their coding. Thus there are problems in finding efficient control designs. The coding operation is also difficult to administer. It is time-consuming and costly and in large-scale operations the coders must sometimes be hired on a temporary basis, and the consequences for maintaining high quality are obvious.

In this paper the importance of good input is stressed. Methods for preventive control and two basic verification procedures, dependent and independent verification, are presented.

Computers have found applications in some survey operations including coding. Considering the various drawbacks associated with manual coding, it appears inevitable to focus on the very basis of manual coding and to investigate the possibilities offered by access to a computer of developing a basically new approach. Here we will review the Swedish efforts concerning this approach called automated coding.

CONTENTS

1 The coding operation and the characteristics of the control problem

Examples of data-processing operations in a survey are editing, coding, key-punching and tabulation. Consider a set of objects ("elements") of some kind and a set of mutually disjoint categories. Each element belongs to one and only one of these categories. Coding denotes the act of assigning the elements into these categories.

In practice coding is based upon access to verbal information about the elements of the population or sample under study. This information is usually obtained on schedules in the data collection operation and is entered either by the respondents themselves or by interviewers or enumerators. Unlike other kinds of information (numerical data on household expenditures, for instance), verbal information cannot be processed immediately into statistical tables. It must first be coded into different categories where each category is labeled with, for instance, a number. These numbers are called code numbers and the key to these code numbers is called the code. (Naturally, numerical data may also be subject to coding. Thus in a census of businesses the objects enumerated can be assigned to categories defined by relevant characteristics, e.g. total turnover).

The term "coding" is admittedly ambiguous. Attempts have been made to replace it by the term "classification"; this term may be better than coding but it has certain disadvantages. Throughout this article the term "coding" is used since it is the one most frequently used in the literature. Some other terms in this area are ambiguous as well. For instance, what in this article is called "code number," is in the literature often referred to as "code," and what here is called "code" is often referred to as "code list," "coding standard," "lexicon," or "nomenclature." Still another ambiguity concerns what is to be coded. In the definition above it was postulated that a given element belonged to a certain category. In the literature coding is often described as an operation in which the verbal descriptions or the responses are coded rather than the elements themselves. This common way of describing the situation is easily understood since in many surveys each element is coded with respect to more than one variable.

The coding operation has three components:

(1) Each element in, for instance, a population is to be coded with respect to a specific variable by means of verbal descriptions.

(2) There exists a code for this variable, i.e. a set of code numbers in which each code number denotes a specific category of the variable under study.

(3) There is a coding function relating (1) and (2), i.e. a set of coding instructions relating verbal descriptions with code numbers.

Coding is a major operation in such statistical studies as censuses of population, censuses of business and labor force surveys. Examples of variables are occupation, industry, education and status.

There are different kinds of coding problems. As with most other survey operations, coding is susceptible to errors. The errors occur because the coding function is not always properly applied and because either the coding function itself or the code is improper. In fact, in some statistical studies coding is the most error-prone operation next to

data collection. For some variables error frequencies at the 10 % level are not unusual. Another problem is that coding is difficult to control. Accurate coding requires a lot of judgement on the part of the coder, and it can be extremely hard to decide on the correct code number. Even experienced coders display a great deal of variation in their coding. Thus there are problems in finding efficient designs for controlling the coding operation. A third problem is that many coding operations are difficult to administer. Coding has a tendency to become time-consuming and costly: for instance, in the 1970 Swedish Census of Population carrying out the coding took more than 300 man-years. In many countries coders in large-scale operations must be hired on a temporary basis, and the consequences for maintaining good quality are obvious. There are even reasons to believe that in the future it might be difficult to obtain even temporary coders for this kind of relatively monotonous work. So there is certainly room for new ideas on the effectiveness of the coding operation.

An overview of the problems with control of coding is given in Lyberg (1981).

2 Broad categories of coding operations

Coding can be carried out manually outside an agency, manually within an agency, or automatically by a computer.

Manual coding outside an agency is more common than one might think. In the continuing Swedish Labor Force Survey the coding of occupation is done by interviewers. In the 1975 Swedish Census of Population, local authorities coded some of the census variables in order to make it possible to produce some employment statistics without the usual time lag. A third example is when the respondents themselves code different variables. At one extreme we have the case when respondents receive a copy of the code and are asked to use it for coding purposes. At the other extreme we have the case called "self-coding" where the respondent is presented with a number of fixed alternatives and asked to choose one of them.

Coding outside an agency shifts the burden of coding. Generally the procedure generates low coding costs for the agency but the agency's control of the coding is reduced as well.

The literature on evaluating of this kind of decentralized coding is not very extensive.

Manual coding inside an agency (centralized coding) is common. At Statistics Sweden such coding is used in over 100 surveys each year. However, in many of these cases, the coding operation is small and presents no serious problems. In other and more interesting cases, we have large-scale coding on a continuing basis. Examples include the censuses, in which Statistics Sweden during the last decades has hired around 300 coders on a temporary basis for each census. The coding operation is complex and many variables must be coded for each element. Such operations put tremendous pressure on the central staff and the organization. Examples of such operations are continuing surveys where variables such as occupation, industry, education and employment status are coded. Such surveys are the Labor Force Survey (both centralized and decentralized coding), different Pupil Surveys, Income Distribution Surveys, etc. The coding in these surveys is done by a regular coding staff.

Since large-scale coding is costly and time-consuming, computer coding could be an attractive alternative. Automated coding is used as a complement to manual coding: manual coders take care of cases rejected by the computer. The methods for automated coding have been used experimentally for some years. A few years ago it was applied for the first time in production, in the 1978 Household Expenditure Survey where the variable "goods" was coded by computer. An immediate successor was the Swedish Author's Fund where authors and book titles were coded. Some important present applications are the 1985 Household Expenditure Survey and the 1985 Census of Population.

These different types of coding operations generate specific problems. In a decentralized coding situation there are problems involved in supervising and controlling the operation. In a centralized situation one often faces complex coding and has to make compromises between quality and cost. In computerized coding one must administer parallel manual systems and be alert with respect to the performance of the computer programs.

3 Coding errors

3.1 The meaning of "coding error"

In this article, we assume that a true code number exists for each element with respect to the variable under study. A coding error occurs if an element is assigned a code number other than the true one. This seemingly simple definition needs some further elaboration. Three different aspects will be considered.

First, it is often difficult to decide upon a true code number. The basic assumption is that every element belongs to one and only one category. In practice there are difficult situations where a specific description is such that it can be assigned different code numbers depending on interpretation.

Second, even if the description is detailed, problems might arise in assigning true code numbers. Who are the experts to decide the code numbers? Studies show that the variation between "experts" can be considerable and as a consequence true code numbers often have to be defined operationally. For instance, three or more experts code the same element. By means of a majority rule a true code number is assigned; i.e., the code number assigned by a majority of the experts is considered the true one.

Third, consider the following example. A dentist fills out a mail questionnaire. One of the questions is "What is your occupation?" For some reason or another the dentist answers "brain surgeon." Thus the information available to the coder is "brain surgeon"; if the code number for "brain surgeon" is, say, 411 and the coder assigns 411 then the coding is correct. If any other code number is assigned, including the one for dentist, a coding error has occurred. Thus we say that the coding operation starts with the available element description, whether or not it is proper, and ends with the assignment of a code number. This limitation is practical from the standpoint of control. The obvious response error in the example must be dealt with by other means.

When the code has two or more digits, the notion of a coding error must take this fact into account. Assume for example that six coders code a specific element with respect to the variable "industry" using a 6-digit code. The outcome may be as follows:

True code number		3	6	9	9	2	2
Coder	1	3	6	9	9	2	1
	2	3	6	9	9	1	1
	3	3	6	9	1	1	0
	4	3	6	2	0	9	0
	5	3	5	5	1	1	0
	6	2	1	0	0	0	0

According to the definition of "coding error" all six code numbers assigned are wrong since none of them coincides with the true one. However, the errors are of different kinds. The code number, 3 6 9 9 2 1, assigned by coder 1, differs from the true one with respect to the last digit. The second coder has been able to code correctly down to the fourth level. The third, fourth och fifth coders have been able to code correctly down to the third, second and first level respectively. The sixth coder has not been able to assign the first digit correctly, and as a result the element has been coded to the main group "mining" instead of "manufacture". The point is that as soon as an error occurs on a specific level all subsequent levels are erroneously coded as well. Furthermore, as a consequence of the way an x-digit code is constructed, an error in the first position is more serious than an error in the second position which, in turn, is more serious than an error in the third, etc. For instance, an error in only the sixth position does not affect the quality of a presentation of results on the fifth level, but an error in the first position affects the presentation of results on any level.

The error rate is the number of incorrectly coded elements divided by the total number of elements coded. The error rate can be calculated for a specific material, for a specific time period, for individual coders and for different levels of the code.

Error rates are gross errors. Normally the results of the coding operation are displayed in statistical tables. The coding errors which remain in the table are net errors and can sometimes be much less than the gross errors, since the errors tend to some extent to cancel out. However, systematic coding errors can seriously distort statistical tables.

There is ample evidence that the coding operation is susceptible to errors. Error rates between 10 and 20 percent when coding variables such as occupation and industry are not uncommon.

Coding variability between and within coders can be substantial as well. The size of the between-coder variability depends heavily on the number of coders involved, but there are examples of within-coder variabilities that exceed 10 percent, for example, we might have a situa-

tion where a specific coder is presented with the same set of elements twice (with an intermission between trials in order to avoid memory effects).

One situation in which coding errors are always serious is the following. Suppose that we want to investigate people with special occupations for a health study. We have at our disposal a census file including all individuals in the area. But we are interested only in miners, stone-cutters and house-painters in our study of, say, pulmonary diseases. By means of the code numbers associated with these three occupations, it should be possible to screen the appropriate population. The drawback associated with incorrect coding now becomes obvious. When investigating the screened subpopulation we notice that some of the people under study do not in fact belong to the three occupation categories. The removal of those elements is both a financial and administrative problem. Much worse is that many miners, stone-cutters, and house-painters are hiding in the rest of the file under false code numbers. These general experiences have occurred on a global scale.

3.2 Some Swedish experiences

3.2.1 An example from a Labor Force Survey

An early Swedish study concerning coding errors (Olofsson (1965)) treated the coding variability for the variables occupation and industry. It was found that coding errors seriously affected the estimation of parameters for gross changes, i.e., the flow between different categories. The main results were

i) only 40 percent of the changes in major (one-digit) occupation categories were real; the rest were coding errors. The corresponding figure for industry was 46 percent.

ii) only 30 percent of the changes in two-digit occupation categories were real. The corresponding figure for industry was 34 percent.

Obviously, publishing such estimates would indeed create an exaggerated picture of mobility in the labor market. In fact, for some categories these coding errors lead to an overestimation of 100-200 percent.

3.2.2 The 1965 Swedish Census of Population

In 1967 an evaluation study of coding errors in the 1965 Swedish Census of Population was conducted (see Lyberg (n.d.) and Dalenius and Lyberg (n.d.)). From a population of census material comprising about 70 percent of the 1965 population, a two-stage sample of verified census schedules was selected. The population was partitioned into four strata and four subsamples were obtained. The evaluation study was confined to the following variables:

(1) Relationship to head of household
(2) Type of employment
(3) Status
(4) Industry

The codes used for variables 1-3 were one-digit-codes; the code used for "industry" was a three-digit-code.

Since we were dealing with four variables and four subsamples, we obtained 16 different estimates of error rates. These are given in Table 1 below.

Table 1. Estimates of error rates (%) in production coding in the 1965 Swedish Census of Population.

Subsample	Variable 1	2	3	4
1	1.6	2.7	1.0	14.5
2	1.4	1.6	.6	8.2
3	1.5	3.0	1.3	14.5
4	.7	1.3	1.2	8.7

Subsamples 1 and 3 consisted of totally verified schedules and subsamples 2 and 4 consisted of sample verified schedules. Most of the total verification was done for still inexperienced production coders; this explains the differences in error rates between total and sampling verification.

In Lyberg (n.d.) studies concerning within-coder replication in the evaluation of the 1965 Swedish Census are presented. Each of the three coders X, Y, and Z in the experiment made one original coding (trial 1). After three weeks the material was coded once again by the same coders (trial 2). These independent trials gave the estimates shown in Table 2 of the within-coder variability $P = m/n$, where n is the total number of coded elements for the specific variable and m is the number of elements differently coded when comparing the two trials.

Table 2. Estimates of within-coder variability, P (%) in the 1965 Swedish Census of Population

Variable	P for coder			Subsample
	X	Y	Z	
Industry	7.4	5.7	4.5	1
	6.5	5.2	3.2	2
	6.7	1.8	1.8	3
	5.8	5.3	2.9	4

The variability illuminates the difficulty of coding the more complex variables.

3.2.3 The 1970 Swedish Census of Population

There were more variables to be coded in the 1970 Swedish Census of Population than in the 1965 Census. For evaluation purposes a sample was drawn from the population of census schedules. A pool of expert coders was used to generate a set of "true" evaluation code numbers for each schedule in this sample. These code numbers were compared with the production code numbers after verification, and this led to estimates of error rates for the different variables on economic activity. These variables were

(1) Relationship to head of household
(2) Type of activity
(3) Occupation
(4) Status
(5) Industry
(6) Place of work
(7) Type of conveyance to place of work
(8) Number of hours at work

Estimates of error rates for these variables are given in Table 3.

Table 3. Estimated error rates in coding economic activity in the 1970 Swedish Census of Population

Variable	Code	Percent error rate (total population)
(1)	1-digit	4.3
(2)	1-digit	4.7
(3)	3-digit	13.5
(4)	1-digit	3.7
(5)	4-digit	9.9
(6)	1-digit	8.9
(7)	1-digit	11.5
(8)	1-digit	4.4

The error rates for the variables (1), (6) and (7) are probably overestimated, since the code numbers were processed by an optical character recognition machine and we have reason to believe that technical errors in this phase had a minor effect on the error rates for those variables.

The table shows that the multi-digit variables are difficult to code. But also the one-digit variables, a priori considered easily coded, are erroneously coded relatively often. One reason could be that the coding

situation is too complex for one coder; i.e. each coder has more variables to manage than he/she can handle.

In the evaluation of the 1970 census coding the experiments for investigating the within-coder variability were repeated. This time five expert coders were used. The results are given in Table 4 below.

Table 4. Estimates of within-coder variability (%) in the 1970 Swedish Census of Population

Variable	Expert coders				
	A	B	C	D	E
Relationship to head of household	.7	1.2	2.4	1.1	.8
Type of activity	1.2	2.1	3.0	1.5	1.8
Occupation	8.0	10.6	10.9	9.2	7.1
Occupational status	2.4	.9	1.8	1.1	1.9
Industry	3.7	8.8	11.6	6.9	5.4
Employment	1.6	3.2	3.9	2.7	2.1
Type of conveyance to place of work	.8	2.7	6.0	1.4	2.9
Place of work	1.3	1.5	2.1	1.8	2.5

The estimates are based on sample sizes ranging from 300 to 1000; this range reflects that the economic variables are coded only for economically active persons. There were also some differences in expert coder workload.

As Table 4 shows the within-coder variability is considerable. This is disturbing when we remember that these coders were used as producers of "true" code numbers to evaluate the coding operation.

3.2.4 The 1975 Swedish Census of Population

The number of variables was smaller in the 1975 Census of Population than in the 1970 Census. Evaluation studies show that the error rates also were smaller in this census than in the 1970 Census. The following variables were studied:

(1) Relationship to head of household
(2) Type of activity
(3) Occupation
(4) Status
(5) Industry
(6) Type of employment
(7) Type of conveyance to place of work

All of these are one-digit variables except for (3) and (5). In Table 5 estimated error rates are given for these variables.

Table 5. Estimated error rates in coding economic activity in the 1975 Swedish Census of Population.

Variable	Code	Percent error rate (total population)
(1)	1-digit	.6
(2)	1-digit	.6
(3)	3-digit	7.8
(4)	1-digit	.5
(5)	4-digit	3.5
(6)	1-digit	1.0
(7)	1-digit	.5

The results given in this table differ strikingly from those obtained in the 1970 evaluation study. The error rates have dropped for every variable and it is most encouraging that the one-digit variables now seem to be much more easily coded. The occupation error rate of almost 8 percent is still serious, but compared to the 13.5 percent rate in 1970, it is a good result. Even better is the estimate for industry.

3.2.5 Some other studies of error rates at Statistics Sweden

Most of the coding studies at Statistics Sweden have been carried out within the censuses. This is rather natural since coding is an extensive operation in a census. During recent years interest in coding errors has grown and as a result, some evaluation studies have been carried out in other surveys as well. Here some estimates of coding errors from such studies are given.

In Olofsson (1976) an industry error rate of 5.7 % is noted in the 1974 Labor Force Survey. Occupation in the same survey had an error rate of 6.2 %. In Harvig (1973b) an 11 % error rate in occupation coding is estimated for coding data for university graduates. In Harvig (1973a) a 3.2 % error rate is estimated when coding underlying causes of death. In Lyberg et al. (1973) an 8 % error rate is estimated when coding

teacher's education. In this case, the 95 % confidence interval was 5.9 % - 10.3 %.

Extensive reviews of studies of error rates in industry and occupation coding are given in Lyberg (1983).

4 Control of coding operations

4.1 The need for control

The general experience seems to be that coding errors do not affect tables of overall statistics very much, since gross errors have a tendency to cancel out and become rather small net errors.

Statistical tables on overall statistics are seldom, however, the single and final output from surveys. Statistics in breakdowns may be seriously in error even if the overall statistics are not. Besides, most surveys are multi-purpose and coded materials are often saved for fu- ture known or unknown analyses. It is common that a coded material with large gross errors is presented as a frequency distribution, say $N_1, N_2, \ldots, N_k$ for k categories, where the net effects of coding errors are small. After some time it is decided that a new survey or a special analysis should be carried out for individuals belonging to one or a few specific categories. At this stage, the gross errors may have serious consequences.

Other difficult situations occur when the material is used in cross-classification or in prediction (as was the case in the labor market mobility example given in Section 3.2.1).

We have now seen why it is so important that coding control is included in the overall program for producing the statistics. However, knowledge of the error rate is not enough if we want to be far-sighted. We need to know about the error structure, the reliability of the coding process, different types of errors, the seriousness of different errors and the effects of errors, in order to take suitable corrective measures with respect to the code or the coder.

4.2 The US Bureau of the Census' survey model

It is beyond the scope of this article to elaborate on the survey model presented in Hansen et al. (1964); we will be satisfied with a brief discussion of the decomposition of the MSE. The population mean $\bar{X}$ is estimated by $\bar{y}$, say, by means of a simple random sample of elements. Then the total error $\bar{y} - \bar{X}$ is measured by $MSE(\bar{y})$, which can be written as:

$$MSE(\bar{y}) = \frac{\sigma_S^2}{n} + \frac{\sigma_R^2}{n}\left[1 + (n-1)\rho_R\right] + \frac{2(n-1)}{n}\sigma_{RS} + B^2.$$

The first term is the sampling variance, the second is the total response variance, the third is the covariance of the response and samp-

ling deviations and the fourth is the squared bias. It is important to remember that the sampling variance measures variations induced by the sampling process, while the response variance measures variations assumed to characterize the measurement operation. An important feature of the model is its broad applicability: it may be applied to any sequence of survey operations, i.e. either the full sequence or a subset of operations (for instance, interviewing and coding). Applied to the full sequence, the response variance reflects contributions from all operations such as interviewing, coding, editing and so forth. Applied to coding alone, for instance, the response variance reflects only coding and the response variance becomes a coding variance. Consequently, coding gives a contribution to MSE of the form

$$\frac{\sigma_C^2}{n}\left[1 + (n-1)\rho_C\right] + B_C^2 \ .$$

The variance consists of a simple coding variance and a correlated coding variance.

In US Bureau of the Census (1972b), Jabine and Tepping (1973) and Bailar and Dalenius (1969) it is described how the variance components of $MSE(\bar{y})$ may be estimated. Essentially the estimation is carried out by means of replications and interpenetrations. Bailar and Dalenius develop a set of basic study schemes in order to obtain observations useful for the estimation process. In Hartley and Rao (1978) a general procedure is provided for estimating the total variance, including coding variance, directly from the current survey data. Naturally, this procedure requires special survey designs involving some kind of imbedded experiments. The Hartley-Rao procedure is improved in Biemer (n.d.) and the latter also incorporates a coder allocation scheme.

Thus the US Bureau of the Census' survey model has two roles with respect to coding control. First, it can help strike an appropriate balance between various control efforts with respect to all survey operations. Second, it enables us to dissect the coding error in a given coding operation.

The emphasis in the studies conducted at Statistics Sweden is in general not explicitly directed towards the measurement of coding errors as they appear in the MSE-relation. Instead, in order to improve coding procedures, we have tried to conduct studies aiming at identifying inaccurate coding procedures, finding efficient verification procedures, and illuminating error structures.

4.3 Schemes for statistical quality control

Manual coding can be characterized as an endless sequence of operations, and it thus seems rather well suited for the application of statistical quality control schemes as originally developed for industrial applications. More specifically, control of coding could be based on various quality control sampling plans. However, coding differs somewhat from, say, car manufacturing. Often there is a problem in finding the true code number, and this forms a sharp contrast to the

situation where the diameter of a screw nut is to be checked. As a consequence, errors of the first and second kind are usually much more common in administrative applications than in other quality control areas. Furthermore, it is often impractical or even impossible to establish risk functions for producers and consumers, since coding is only one part of the statistical production process. Nevertheless statistical quality control has been used for several decades as a means for keeping the desired quality level of coding.

A sampling inspection plan can assure quality of any prespecified level. The literature discussing such plans is extensive; an early and well-known example is Dodge and Romig (1944).

The statistical quality control of coding aims at controlling the code numbers assigned (a control for the producer of data). The major instruments available are acceptance sampling, process control, and combined procedures which utilize both acceptance sampling and process control. Applications of these techniques may be found in Fasteau et al. (1964), Minton (1969), US Bureau of the Census (1965) and Minton (1970b).

A common approach is to use a combination or hybrid of acceptance sampling and process control when dealing with administrative applications such as coding.

4.4 Specific coding control schemes

There are certain control schemes designed specifically for coding. These schemes are applicable in three different areas, namely

- Training of coders
- Dependent and independent verification
- Evaluation

4.4.1 Training of coders

The training and education of coders is indeed valuable since the error rate often decreases with time. If it is possible to "cut" error rates at the beginning of a coding process, one will probably end up with a higher average outgoing quality.

Literature on the training and education of clerks is not very extensive. However, the subject is discussed in Minton (1969) and in Dalenius and Frank (1968).

4.4.2 Verification

There are two main schemes for verification of coding, i.e. for deciding whether or not a code number is correctly assigned: dependent and independent verification. In dependent verification the verifier has access to the code number assigned by the production coder. In independent verification the verifier has no such access and the decision on outgoing code number must be based on different rules such as majority or modal rules. Dependent and independent verification is dealt with in Lyberg (n.d.), Lyberg (1969) and Minton (1969).

Let us start with a definition of dependent verification.

> An element is coded by production coder A. The code number is then reviewed by verifier B. B inspects the code number assigned by A and decides if it is correct or not. If it is considered correct it remains unchanged; otherwise B changes the code number.

With this type of verification experience tells us that the verifier has a tendency to let erroneous code numbers remain unchanged: his judgment is influenced by (depends on) the code number assigned by A. Various studies show that the proportion of incorrect code numbers that remain unchanged can be substantial. At Statistics Sweden we often use 50 % as a rule of thumb.

Since the disadvantages of dependent verification can be traced to the element of dependence, the obvious option is some kind of independent verification. This latter alternative has been used, for instance, in the 1960 and 1970 US Censuses of Population and Housing and in the 1970 and 1975 Swedish Censuses of Population and Housing.

Independent verification is defined as follows:

> An element is production-coded by a coder A. The code number is denoted x_A. The element is also coded by N other coders, where $N \geq 1$. Their code numbers are denoted $x_B, x_C, \ldots, x_{N+1}$. Each coder works independently and as a consequence does not know how the others have coded. Thus we end up with a set of code numbers $x_A, x_B, \ldots, x_{N+1}$. These code numbers are matched and a decision rule defines the outgoing code number.

This definition gives rise to at least two questions. How do we create a situation in which each coder works independently of the other coders? What decision rules are possible?

The first question is a matter of administrative resources. One option is to code directly on the schedule and then mask the code numbers after each coding. This however, is a rather clumsy procedure, and is seldom used. Another option is to copy the schedules to be verified. This procedure is costly, of course, but it has been used in some studies. A third alternative is to code on special forms. This is the best alternative so far, and could work smoothly in a computerized environment, where the matching is done by a computer program.

There are various possible decision rules. One rather natural one is the majority rule: if a majority of the N+1 coders involved agrees upon a specific code number, then this code number is the outgoing one. (If a majority is not reached, then special measures are taken.) One early example of this rule is the use of the three-way independent verification system in the coding process of the 1960 US Censuses of Population and Housing. With three coders involved in each decision, we have three possible outcomes, which can be denoted 3-0, 2-1 and 1-1-1; a majority is reached except for the last case.

A very natural way to improve the efficiency is to use a sequential procedure. In the three-way system this means that we start with two coders. After the matching of their code numbers, it is decided whether a third coder is needed or not. Obviously if the outcome of the first

matching is 2-0, then the code number of the third coder is completely unnecessary to reach a majority. His/her contribution could only lead to either 3-0 or 2-1, and we have probably wasted some money. However, if the outcome is 1-1 the third coder must enter the scene and his/her contribution leads to a majority 2-1 case or a 1-1-1 case. This sequential system, sometimes called two-way independent system, was used in the 1970 US and 1970 Swedish Censuses of Population and Housing.

Independent verification can be carried out in many different ways. It is, almost by definition, more reliable than the various types of dependent verification. However, it is costly and time-consuming, and this makes its introduction difficult.

Verification schemes can be administered on a total or on a sampling basis. Several problems occur when using sampling inspection plans in the verification operation, although using such plans is a natural way of solving the budget problem. Often a rectifying scheme is used and it is necessary to let coders flow between total control and sampling control; i.e., it is necessary to control both the product and the coders dealing with it. A great problem in production control is the fact that inspection is not error-free. The impact of these errors on single sampling plans is discussed in Minton (1972). The flow must be regulated by means of some prespecified criterion. In Cook (1959) a special point system is presented in which each coder receives a point for each erroneous coding, and decisions about the fate of the coder are based upon the accumulated number of points received when coding a specified number of schedules. In Minton (1970a) some other decision rules for administrative applications of quality control are discussed.

4.4.3 Evaluation

Evaluation of coding results provides a basis for the allocation of quality control efforts. We have already given examples of results from different evaluation studies. The results of such studies give suggestions concerning the size and emphasis of the necessary quality control program.

Evaluation studies assume the existence of "true" code numbers, which usually are those generated by more skilled coders or expert coders. By comparing these true code numbers to those assigned by the production coders an estimate of the gross error rate for the production coding can be obtained.

It is obvious that many of the coding errors do not depend on the ability of coders: the codes and the coding manuals may be insufficient and thus cause great variability in the coding process. Improvements of these tools therefore seems an urgent task. Possible action consists of merging categories, making the code less complex, and bringing the definitions of the categories closer to reality.

5 Preventive control

5.1 The importance of "good input"

One of the axioms in quality control is that you cannot inspect quality into a material: the quality must be there from the start. If produc-

tion coding is of substandard quality prior to verification and control, then the costs of verification and control will be excessive. Let us illustrate the effects of "good input" with a few examples.

5.1.1 Choice of code

If it is possible for the designer of the coding operation to influence the construction of the code, he/she should certainly take this opportunity, since it can improve the outgoing quality. During the 1970 Swedish Census of Population it was found that some of the verbal descriptions of occupation tended to be erroneously coded because the code was so detailed that it did not fit some of the rather vague descriptions given. As a consequence, in the 1975 Census some categories of occupation were merged into broader categories. Examples of such merging were

i) chemists + physicists
ii) tailors + dressmakers
iii) housepainters + lacquering men
iv) telephone operators working at telephone company + telephone operators working at offices.

The code changes resulted in less general vagueness and as a consequence we achieved a smoother control operation. Of course this was achieved with the loss of information about the previous, more narrow, categories. However, since these were considered unrealistic we could certainly live with that loss.

5.1.2 Verification of "good input"

It is typical of dependent verification that the verifier does not identify all errors made by the production coders in the inspected material; the proportion of errors identified will depend on the skill of the verifiers, as does of course, the proportion of errors made by the production coder. This brings up an important question: who is to serve as production coder and who is to work as verifier? We will respond to this question by means of an example.

Consider a situation with two categories of coders; the number of coders is the same in both categories. One category codes with a 10 % error rate and the other with 20 %.

Let us assume that - irrespective of coder category - the proportion of errors identified in the course of the verification is stable around the value $\Pi = 50$ %. The usual system is that the "20 %-category" is used as production coders and the "10 %-category" as verifiers. If all the material is verified, the outgoing error rate will be $20 \cdot 50 \cdot 10^{-2}$% or 10 % with the usual system. If instead the "10 %-category" is used as production coders and the "20 %-category" is used as verifiers, the outgoing error rate will be $10 \cdot 50 \cdot 10^{-2}$ % or 5 %. The example will illuminate the idea.

Using the best coders as production coders was suggested in planning both the 1965 and the 1970 Swedish Censuses of Population and Housing but there seems to be a psychological resistance among management for

dealing with the error problem in this way. The evaluation studies show that Π in fact is rather stable, so probably such a system is very useful.

5.1.3 "Good input" in statistical quality control

The efficient choice of an acceptance sampling plan assumes that the error rate before verification, p, is known in advance. If p is high and the desired quality is lower than p, statistical quality control simply does not work; too much rectifying inspection is needed and the result might even be that total verification is a cheaper alternative.

5.2 More on the design of the code

Often the users of survey data want a highly detailed coding that is difficult if not impossible to implement given a specific response pattern of verbal descriptions. In Sweden, at least, there seems to be little if any connection between the empirical verbal response pattern and the code construction. The result is coding variability.

When constructing the code, the following aspects should be included.

The general desiderata should be listed and evaluated simultaneously by the users and producers of the statistics. The desidarata should be weighed against the prevailing response pattern. Sometimes a formal definition of a category gives us trouble: for instance, the user might be interested in a certain partitioning of a broader class of occupations into more narrow categories. However, the response pattern might be so standardized that a coding according to the user's special wishes is impossible. To achieve a correct coding the coder would have to rely on auxiliary information which is perhaps diffuse or simply does not exist. This difficulty seems to be one of the major, if not the major, sources of coding errors. From a dogmatic point of view this kind of error is easily removed: a much better coding can be obtained if special "trash categories" for "non-specified" cases are used whenever there is any doubt about the coding of a specific description. However, we cannot ignore the user's needs and act in this manner all the time. The result would be statistical tables of minor value. Also, the cell frequencies for the "trash categories" would be relatively high, and the result of such a coding would be that we would have in effect a non-response situation. Of course, one solution is to use, in the first place, a better and more costly measurement procedure that makes more accurate information possible, but then we have another survey.

Many of the errors in the 1970 Swedish census coding could have been avoided if code numbers for "non-specified" cases had been used more often. For instance, every major occupation and industry group include a "trash category." These categories were, however, not always used in cases of vagueness, since the administrators were afraid of the "non-specified" problem within major groups. Thus the "non-specified" problem was traded against a coding problem. When designing the 1975 Census it was decided that the "trash categories" should be used more frequently. In order to avoid the "non-specified" problem, or at least reduce it, the regular coding was followed by a special procedure in

which complete information about vague cases was gathered on a sampling basis. These two sources of information were combined, and this resulted in a more valid estimate of minor group totals. The procedure is described in Andersson (1974).

In an evaluation study of the Swedish Labor Force Survey, the main causes of errors were noted for each of the erroneous industry code numbers discovered in the sample. It is quite clear that most of the errors were caused by the code or by the aids used by the coder when applying the code. Only about 24 % of the coding errors could be totally attributed to the coder. This information has lead to several changes and clarifications in the instructions for coders and interviewers.

To recapitulate it is important that "trash categories" be used more often and that the respondent patterns be carefully analyzed for the purpose of collapsing categories. A relatively small number of such amalgamations can reduce the error rate substantially.

5.3 Selection and training of production coders and verifiers

5.3.1 The concept of a master set

The selection of coders and verifiers should be based on coding performance. Information on coding performance is arrived at after the necessary initial education and training period. Both non-production training and test coding can be carried out by means of a master set. The concept of a master set is fully discussed in Dalenius and Frank (1968). The main feature of such a set is that it is known which category each element in the set belongs to.

By using the master set technique the categories themselves can be tested and ill-defined categories can be properly redefined. Another result possible is the screening of coders into different groups - for instance, coders ready for production coding and coders who need more training.

Besides making it possible to identify good coders, the master set can be used during an ongoing production to see whether the coder's performance is deteriorating.

Such a training technique can be run parallel with the ordinary production and its control system. For instance, the ordinary production control system may not be quite adequate to identify bad coders; in this case we can talk about preventive control too.

As pointed out in Dalenius and Frank (1968) and Minton (1968), there are certain practical problems connected with the construction and the use of a master set. For instance, the elements of the set might be chosen from the ongoing production, from certain pilot surveys or from an earlier statistical study of the same general kind as the ongoing production. The last approach seems the most practical, but on the other hand this choice brings with it certain limitations from the user's standpoint. Among other things, it could not serve as a device for process control. However, for the parallel system mentioned above it could serve well.

In the 1970 Swedish Census of Population a master set was constructed to provide a means for efficient preventive control. It turned out, however, that it worked best from an educational point of view and it was used for two weeks before the coders were put on production work.

6 Production control

6.1 An evaluation of dependent verification

In 1967 a study, presented in Lyberg (n.d.), was conducted in order to illuminate, in a concrete fashion, the performance of the dependent verification used in the 1965 Swedish Census of Population. A general evaluation of the coding was obtained as a byproduct.

Until that time the use of dependent verification had been more or less taken for granted. For instance, in the design of the coding control operations of the 1965 Census, no alternatives were proposed. However, evaluation results from the US Bureau of the Census eventually reached us and the dependent verification system was questioned.

In our Swedish study, verified schedules from the 1965 census material were sampled from four strata covering about 70 percent of the population. The stratification was based on geographic area and on whether the schedules were verified totally or on a sampling basis. The study was confined to the following four variables:

(1) Relationship to head of household	(1-digit)
(2) Employment	(1-digit)
(3) Status	(1-digit)
(4) Industry	(3-digit)

The samples were coded by a team of three experimental coders who were considered especially skilled. Each coder coded independently of the others and then the code numbers were matched. Three cases were possible. First, all three coders could agree; we call that case 3-0. Second, two could agree but not the third; we call that case 2-1. Finally, we have the case when no two coders agree; we call that case 1-1-1. In the first and second cases, clearly, we were able to define a majority code number; that number was used as an evaluation code number. In the 1-1-1 cases we let an expert decide the evaluation code number. After that the evaluation code number was compared to both the unverified production code number (P) and the dependent verification code number (V). Table 6 shows the results for variable (1) for one of the subsamples.

Table 6 a. A comparison between the majority code number and the production code number (P_1) (number of cases).

Variable 1, Subsample 1.

Experimental coder combinations	P_1 agrees with ... experimental coders 3	2	1	0	Expert cases	Total
3 - 0	1131	-	-	16	-	1147
2 - 1	-	2	3	0	-	5
1 - 1 - 1	-	-	-	-	0	0
						1152

Table 6 b. A comparison between the majority code number and the dependent verification code number (V_1) (number of cases).

Variable 1, Subsample 1

Experimental coder combinations	V_1 agrees with ... experimental coders 3	2	1	0	Expert cases	Total
3 - 0	1141	-	-	6	-	1147
2 - 1	-	2	3	0	-	5
1 - 1 - 1	-		-	-	0	0
						1152

Obviously the variable is an easy one to code; among the experimental coders only 5 out of 1152 cases caused some kind of disagreement. The error rate among production coders is estimated by summing the number of deviations from the majority code number, i.e. the frequencies inside the triangle. Thus the estimate of the error rate in this case is 19/1147 or 1.6 per cent.

The 19 cases within the triangle were reduced to 9 by means of the verification system and the system handled an estimated 53 per cent of the errors.

Now let us turn to the more complex three-digit variable (4). The corresponding results are given in Table 7 below.

Table 7 a A comparison between the majority code number and the production code number (P_4) (number of cases).

Variable 4, Subsample 1

Experimental coder combinations	P_4 agrees with ... experimental coders 3	2	1	0	Expert cases	Total
3 - 0	427	-	-	48	-	475
2 - 1	-	41	24	8	-	73
1 - 1 - 1	-	-	-	-	5	5
						553

Table 7 b.A comparison between the majority code number and the dependent verification code number (V_4) (number of cases).

Variable 4, Subsample 1

Experimental coder combinations	V_4 agrees with ... experimental coders 3	2	1	0	Expert cases	Total
3 - 0	451	-	-	24	-	475
2 - 1	-	44	23	6	-	73
1 - 1 - 1	-	-	-	-	5	5
						553

This is a much more difficult variable for the coders: they agreed in only 475 of the 553 cases or in 86 per cent. The error rate in production coding, estimated by (48 + 8 + 24)/548, is a striking 14.6 percent.

The 80 cases within the triangle in Table 7 a were reduced to 53 by means of the verification system. This time the system took care of only 34 percent of the errors.

A summary of the error reduction rates for all variables and subsamples in the study is given in Table 8 below.

Table 8. Estimates of error reduction rates (%) when using dependent verification (rounded figures). (Absolute error rate before verification within brackets.)

Subsample \ Variable	1	2	3	4
1	53(19)	16(31)	18(11)	34(80)
2	38(13)	7(15)	33(6)	11(37)
3	32(19)	45(38)	6(17)	30(74)
4	0(9)	41(17)	25(16)	19(52)

Clearly, this type of calculation needs some further elaboration. First, the choice of the majority code number for evaluation purposes certainly has its drawbacks: the 2-1 cases are an indication of variability in the process. To charge a coder with an error when he/she has assigned a code number that agrees with the one assigned by one of the experimental coders is a dubious procedure. In the original study these doubts were handled so that the comparison between production code number or verification code number with the majority code number led to a special classification of disagreements. If the code numbers P or V agreed with none of the experimental coders in the 3-0 case, this was called a "very serious" disagreement. If P or V agreed with none of the experimental coders in the 2-1 case, this was called a "serious disagreement". If P and V agreed with one of the experimental coders in the 2-1 case, this was called a "less serious" disagreement.

As a result using this operational rule might lead to an overestimation of the overall error rate per variable. Furthermore, since the coder is charged with errors resulting from "less serious" disagreements, we probably get an exaggerated picture of the ineffectiveness of dependent verification.

However, the most important observation is that the estimated error reduction frequencies do not come close to the ideal 100 %.

Considering that the cost of the verification operation during the 1965 Census coding was more than 20 per cent of the primary coding cost, some might argue that the whole verification operation was a waste of financial resources. Of course this is not true; the very existence of a quality control program has a salutary psychological effect on all persons involved in the coding operation.

As pointed out in Linebarger et al. (1976), dependent verification has certain advantages. The operation is quick, fairly non-disruptive to handle, requires little work if handled clerically, and is rather inexpensive. The serious disadvantages are to be found in the quality. We must have a reasonably low value of the probability that an erroneous production code number is not changed by the verifier (or changed to another erroneous code number)(β). The tables in this chapter give estimates of β which do not favor the dependent system. And if the disadvantages of a verification system are found in the quality, we are in big trouble.

6.2 A note on independent verification

Today at Statistics Sweden independent verification has more or less replaced dependent verification. Independent verification is a more reliable process than dependent verification: it generally allows more accurate estimation of errors and the coding results are more credible. However, the advantages of dependent verification, listed in Linebarger et al. (1976), simultaneously constitute the drawbacks of independent verification. Thus the latter is more expensive, more disruptive to handle, and more time-consuming. Of course the cost is important (that was the reason why independent verification was not used throughout the entire 1970 Swedish Census of Population). There is also a certain delay in feedback operations with independent verification. For instance, there is a tendency not to discover deteriorating coder performance until rather late, and correction measures are thus delayed. On the other hand, more errors are detected and the picture of the error structure is clearer with independent verification.

The basic idea in all independent schemes is that the outgoing code number (majority or modal) is very likely to be the correct one. Seldom will a minority coder be unjustly charged with an error. However, it is important that the coders whose code number are compared be of approximately equal ability. In Harris (1974), it is pointed out that two "poor" coders could overrule a "good" coder simply because the former have not read the instructions properly. Harris gives an example from Mortality Medical Coding which is operated under a sequence of coding instructions. The example goes like this.

> If diagnosis X is listed, code 111. A "poor" coder might stop here, assign 111 and go on to the next coding unit. However, there may be an additional instruction which says whenever code number 111 appears, check to see if diagnosis Y is also listed. If yes, change the code number to 123. The "good" coder often arrives at 123 and as a "reward" he/she is charged with an error.

The hypothesis that "poor" coders adversely affect "good" coders has been tested by the staff at the US Department of Health Education, and Welfare. The hypothesis was rejected. However, it is obvious that this case must occur now and then; i.e., the probability of making the same error could be substantial. For instance, during a Census of Population additional coding instructions are produced continuosly (at least in Sweden) or at least as soon as "new cases" occur. Obviously there is a risk that less ambitious coders might fail to learn the new instruction quickly and might as a result overrule a more dutiful coder.

In Boston (1977) some doubts are thrown on the basic idea of independent verification. An expert evaluation of cases of coding agreement shows that the error rate in occupation coding cannot be neglected: a series of tests showed that the majority code number was incorrect in .7 to 1.8 per cent of the cases. However, everything depends on what is meant by "highly likely to be correct." Compared to the error rates obtained with dependent verification, we are usually still in a favorable position with independent verification. Besides, in this series of tests only one expert was used to determine if the agreement cases are correct or not. Experience shows that there are no experts in complex coding (by "experts" we mean persons with unit probability of correct coding). This lack of experts creates both theoretical and practical problems.

7 Automated coding

7.1 The challenge

As illustrated, manual coding has various drawbacks. Especially it is: time-consuming, costly, error-prone and boring.

To cope with these drawbacks, it appears inevitable to focus on the very basis of manual coding and to consider the possibilities offered by access to a computer for developing a new approach. This idea is, of course, not in principle new; for instance, at the US Bureau of the Census geographic coding has been conducted by means of computer since 1963. What is new is the suggestion in that agency that the computer be used extensively in the coding of such complex variables as occupation and industry. This suggestion may be viewed as a natural extension of earlier uses of computers in the editing operations.

During the last decade we have conducted a series of experiments in Statistics Sweden in order to find out whether or not it is possible to automate the coding process.

7.2 A bird's-eye view of automated coding

In automated coding we distinguish four operations:

i) Construction of a computer-stored dictionary;
ii) Entering element descriptions into the computer:
iii) Matching and coding;
iv) Evaluation.

7.2.1 Construction of a computer-stored dictionary

In automated coding a dictionary stored in the computer takes the place of the coding instructions used in manual coding. Obviously the construction of such a dictionary is an important task. The construction work could be carried out manually but, when dealing with multi-digit variables, using the computer seems to be a better alternative. The resulting dictionary should consist of a number of verbal descriptions with associated code numbers. The descriptions could be a sample from the population to be coded or a sample from an earlier survey of the same kind. Of course an important problem is the size of the sample underlying the dictionary construction. Whether the dictionary is constructed manually or by computer, its code numbers should be those assigned by the best of the available coders. We should also use independent verification procedures.

7.2.2 Entering element descriptions into the computer

Verbal descriptions are to be entered into the computer. One possible method is to punch the descriptions in a more or less free format on cards or magnetic tape. However, this method has some serious drawbacks: first it consumes a lot of "space," and second, the errors involved in large-scale keypunching of alphabetic information are relatively unknown; moreover such keypunching is rather costly.

A better alternative would be to have the verbal information directly available for optical character recognition. Unfortunately the recogni-

tion of hand-written letters is not yet sufficiently developed for this purpose.

At the present, there are reasons to believe that the entering of verbal descriptions to the computer is the most important practical problem in designing systems for automated coding.

7.2.3 Matching and coding

Each element description put into the computer is compared with the list of occupation descriptions in the dictionary. If an element description agrees with an occupation description (is a "match"), it is assigned the corresponding code number; otherwise it is referred to manual coding.

In an automated coding system we will obtain exact matching for a fraction of all elements only. A primary task in developing such a system is to design criteria for the degree of similarity between input words and dictionary words necessary for them to be considered a match.

7.2.4 Evaluation

The system must include continuing evaluation studies. Such studies aim at

i) controlling the quality of computerized coding;
ii) improving the dictionary and;
iii) controlling the cost.

Whether automated coding is profitable or not is a question to be answered by the evaluation. Are the referred cases more difficult to code than those taken care of by the computer? Does the dictionary need improvement? These and other questions are to be resolved by evaluation.

7.3 The dictionary

There are two general kinds of algorithms for automated coding: weighting algorithms and dictionary algorithms. Weighting algorithms assign weights to each word-code combination using information from a basic file. When a new record is to be coded, the program chooses the code number which is assigned the highest weight for the specific record word. Dictionary algorithms look in a dictionary for words or word strings that imply specific code numbers. When a new record is to be coded the program determines whether the word or word string matches any word in the dictionary. If no match occurs, the record is rejected and referred to manual coding.

At the US Bureau of the Census a number of different algorithms have been developed and investigated during the last decades. In some straightforward applications, like the geographic, coding automated coding has been quite successful. Recent efforts deal mainly with the more complex coding of occupation and industry. Four algorithms are described in Lakatos (1977a, b). Two of them, the O'Reagan and the Corbett algorithms, use dictionary methods. The remaining two, the IMP and the INT algorithms, use the weighting method. The INT algorithm was developed by Rodger Knaus, and was further described in Knaus (n.d., 1978a, b, 1979, 1983). Current development work at the US Bureau of the Census is

described in Appel and Hellerman (1983).

At Statistics Sweden we have worked with the dictionary approach only. Thus we have nothing to add with respect to other algorithms.

The computer-stored dictionary is a parallel to the dictionary and the coding instructions used in manual coding. In order to create such a dictionary a number of operations must be carried out:

i) Choice of basic material;
ii) Sampling a basic file from the basic material;
iii) Expert coding of the basic file;
iv) Establishing inclusion criteria for dictionary records;
v) Construction of a preliminary dictionary;
vi) Testing and completing the preliminary dictionary.

A meaningful description of these operations would demand considerable space. The interested reader is referred to Lyberg (1981) for details.

7.4 Computerized construction at Statistics Sweden

Our present dictionary construction system at Statistics Sweden generates a dictionary with two chapters, PLEX and SLEX. PLEX contains unequivocal descriptions and is scanned first. SLEX contains discriminating word strings that fit several different input descriptions. As a consequence SLEX is not as accurate as PLEX and it is scanned only if PLEX fails to code. Our experience shows that it is rather easy to construct a PLEX manually, but that manual SLEX construction is much harder to manage. We have made a program for computerized construction of SLEX. (As a consequence a computerized PLEX is obtained as a simple special case.)

We have tried a few different versions of the program. The present version, a package called AUTOCOD, is described in Bäcklund (1978). All programs are written in PL1. AUTOCOD contains routines for

- the creation of computer stored dictionaries (PCLEXK)
- the coding of descriptions (PCAUTOK)
- the updating of dictionaries (PCLEXUP)
- the evaluation of dictionaries (PCLEXT)

PCLEXK creates a PLEX and a SLEX. The procedure involves three steps. The program LEXLADD creates space for a possible SLEX. LEXKONS creates PLEX and SLEX. For each PLEX description, say, a six-character abbreviation starting with the first character is tested for inclusion in SLEX. If that abbreviation fits another PLEX description, it is rejected and a new abbreviation is created starting with the second character of the PLEX description. The procedure is repeated at most six times; if no valid abbreviation is obtained the procedure goes on to the next PLEX description. Finally LEXLIST lists the dictionaries by means of EASYLIST. Parameters that can be varied include:

- possible use of a list of prefixes which, when making a dictionary of, say, goods, removes such word strings as pounds, roll, and pairs
- minimum frequency f_0 (the dictionary inclusion criterion)
- tolerated degree of equivocalness
- minimum length of words in SLEX.

PCAUTOK codes new records by means of PLEX and SLEX. PCLEXUP is used when we want descriptions to be removed from or added to an existing dictionary. PCLEXT is used to evaluate a dictionary when we have access to a material with manual code numbers assigned.

PLEX and SLEX can be updated simultaneously or separately.

7.5 An example of an application

Automated coding has been applied in some regular productions at Statistics Sweden. The very first application was the coding of goods in the 1978 Household Expenditure Survey. After that automated coding has been applied in coding occupation and socio-economic classification (SEI) in the 1980 Census of Population and in coding goods the 1985 Household Expenditure Survey. These are the major efforts so far. Automated coding is also used in a minor continuing survey of book loans where authors and book titles are coded. It is also used in coding occupation and SEI in the continuing Survey of Living Conditions and in coding occupation in Pupil Surveys. In this section we will describe one of these applications, namely the coding of goods in the 1978 Household Expenditure Survey (HES).

In the 1978 HES approximately 5900 households were supposed to keep a complete diary (CD) of all goods purchased during a two-week period. The rest of the sample, approximately 7900 households, was supposed to keep a simplified diary (SD) of goods purchased during a four-week period.

The survey design allowed continuous delivery of diaries from the respondents. The material could be processed in cycles, which might be advantageous in a system with automated coding. Since we were not at all convinced that our system should work in a production environment, it was decided that, to start with, the coding should be carried out by two parallel systems, one manual and one automated. After two months production the systems should be evaluated in order to decide a preferred system to be used during the remaining ten months of production.

The dictionary construction was step-wise. Extensive efforts were made in creating an initial dictionary. After that continuous revisions were made prior to many cycles. Each unique description was coded by HES experts. Only a PLEX was constructed, with a 100 % unequivocal rate. This initial dictionary consisted of 1459 descriptions. In the automated coding procedure, uncoded descriptions were listed alphabetically on an optical character recognition form and code numbers were assigned directly on it. Some of the uncoded descriptions were added to the dictionary in the updating process.

During the survey period 33 cycles were run. During this period 17 different versions of PLEX were used; only a few cycles were coded with identical dictionaries. In Table 9 below the dictionary sizes and coding degree for the cycles are given.

The coding degree over all cycles was 65 %. As can be seen from the table, the coding degree decreases sharply now and then. This is explained by the fact that CD's are easier to code automatically compared with the SD's and that the proportion of CD's varies between the cycles.

Table 9. Dictionary size and coding degree for the 33 cycles in the 1978 HES

Dictionary version	Cycles	Number of dictionary descriptions	Coding degree (%)
1	1	1459	56
2	2	1554	63
3	3	1760	67
4	4	2228	66
5	5,6,7	2464	68,68,63
6	8	1632	64
7	9	1990	53
8	10,11	2451	69,66
9	12	2866	61
10	13	3065	68
11	14	3613	58
12	15,16	3752	72,73
13	17,18	3832	39,70
14	19,20	4011	65,73
15	21,22	4229	51,72
16	23,24,25,26,27 28,29,30	4230	64,67,62,67,72,65, 67,50
17	31,32,33		65,39,67

In coding the 1978 HES, it was decided to use PLEX only because of the inefficiency of the SLEX file. The price paid is the lower coding degree: we assume that it goes down 10-15 per cent when SLEX is dropped. At the same time, though, coding quality is high with an error rate, for the 65 % coded, of less than 1 %. This rate can by no means compete with the one a SLEX would give. In all, the coding of the 1978 HES went smoothly. The key operators found it less boring to key in verbal information for a change. The cost calculations point out that automated coding was 2-5 % cheaper that a conventional manual system. Besides, the system provided some further advantages. Since all descriptions are key-punched, the primary material is better documented than when merely the code number is keyed. Consequently it is possible to give more detailed descriptions of the goods contained in the groups for which estimates are provided. Furthermore, since the dictionary manages to code most straightforward descriptions, the remaining manual coding becomes more interesting to the coder.

The computer coding itself is cheap. For instance, coding 7166 purchases in cycles 12 cost 226 SEK (approximately 30 U.S. D). Coding 34070 purchases (extending the volume almost five times) cost 327 SEK. The cost of updating dictionary number 13, for instance, was 366 SEK. Instead, we noted that the manual preparation of runs was the expensive part. Consequently we eventually learned that if the number of dictionary modifications and the number of cycles could be reduced, this would also produce considerable cost reductions.

It does not seem worth the effort to make extensive dictionary revisions after a specific point. Quite soon a rather stable coding degree is obtained which cannot be substantially altered without changing the dictionary construction principle. We note that with the third version

already we have obtained a coding degree of 67 %. Despite much work and repetitive modifications after that point, we have at best obtained 73 %.

8 The future of coding control

The future of coding control, like that of other fields, is hard to predict. However, we know a few things for sure. We know that administrative records will be increasingly used for statistical purposes. In recent years a more intensive use of existing administrative records as sources of data for statistical purposes has been called for, at least in the U.S.A. and Sweden. There are many reasons for this demand. Among other things, the very existence of administrative records calls for an extended use for statistical purposes. Furthermore the respondent burden, heavy as it already is, must not be allowed to increase.

There are reasons to believe that in some cases respondents are more willing to report to an administrative system than respond to a statistical survey. In those cases it seems obvious that a statistical system should coordinate its data collection with a suitable administrative system. Such coordination presents problems of its own, e.g., in differences as to definitions, data collection procedures, etc. Furthermore, the legal issues are far from resolved.

An efficient link between administrative records and statistics production implies the necessity of uniform coding. Uniform coding means uniform codes; i.e., for the variable under study a specific verbal description should get the same code number irrespective of the system. As one might notice, this is not always the case even within statistics production. Uniform coding is the necessary basis for effective coding control and is therefore one of the most urgent tasks.

Coding within administrative systems is of course very sensitive to errors since each error can have important effects on the individuals for whom the system is designed. The coding control must therefore be tighter than in statistical systems, where the errors sometimes tend to cancel eachother out. However, special care must be taken in using administrative systems in statistics production. Suppose that data in an administrative system are used to construct sampling frames for statistics production. Then, again, each individual error might constitute a serious flaw.

Accordingly it seems obvious that one can expect in the future more intensive use of efficient control methods such as independent verification.

The development of computers has been important over the years. We have seen, for instance, that the replacement of manual matching of code numbers by computer matching facilitates things a great deal. Recent studies show that data processing can be made more efficient by letting a single clerk carry out several processing tasks more or less simultaneously. For instance, it has been shown how interviewing can be computer-assisted by means of a system in which telephone interviewing is carried out at a computer terminal. No wild fantasy is needed to imagine that such a procedure can be extended to cover operations such as editing, coding and punching as well. Concentration of several operations, including coding, in a centralized facility makes things a lot

easier. Coding instructions can be displayed on terminals and quick changes in codes and instructions can be entered on terminals. Verification and feedback can be conducted more easily in a computerized environment. Coding in such an environment should be a cheaper and more effective operation.

The success of automated coding is a function of language complexity. It seems that the Swedish language is more forgiving than English in this respect. Our studies have shown that automated coding might be a possible option when designing the coding operation.

9 References

Andersson, R. (1974): Skattning av totaler med utnyttjande av supplementär information. Metodinformation nr 74:14, Statistics Sweden (In Swedish).

Appel, M.V. and Hellerman, E. (1983): Census Bureau Experience With Automated Industry and Occupation Coding. American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 32-40.

Bailar, B.A. and Dalenius, T. (1969): Estimating the Response Variance Components of the US Bureau of the Census' Survey Model. Sankhyā, Series B, vol 31, Parts 3 & 4, pp. 341-360.

Biemer, P.P. (n.d.): An Improved Procedure for Estimating the Components of Response Variance in Complex Surveys. Unpublished memo.

Boston, G.F.P. (1977): Quality Control of Occupation Coding. Unpublished draft, Office of Population Censuses and Surveys, U.K.

Bäcklund, S. (1978): Automatisk kodning. Beskrivning av programvara och programvaruhantering. Memo, Statistics Sweden (In Swedish).

Cook, W.H. (1959): A Point System of Quality Control. In: American Society for Quality Control. Science in Management. Proceedings of the 1959 conference of the Administrative Applications Division, Washington, DC, pp. 207-209.

Dalenius, T. and Frank, O. (1968): Control of Classification. Review of the International Statistical Institute, vol 36:3, pp. 279-295.

Dalenius, T. and Lyberg, L. (n.d.): An Experimental Comparison of Dependent and Independent Verification of Coding. Memo from Tore Dalenius to Leon Pritzker.

Dodge, H.F. and Romig, H.G. (1944): Sampling Inspection Tables. Wiley.

Fasteau, H.H., Ingram, J.J., and Minton, G. (1964): Control of Quality of Coding in the 1960 Censuses. Journal of the American Statistical Association, vol. 59, no 305, pp. 120-132.

Hansen, M.H., Hurwitz, W.N., and Pritzker, L. (1964): The Estimation and Interpretation of Gross Differences and the Simple Response Variance. In 'Contributions to Statistics', presented to Professor P.C. Mahalanobis on the occasion of his 70th birthday, pp. 111-136.

Harris, K. (1974): Analysis of the Independent Three-Way Verification System in Mortality Medical Coding. Memo, US Department of Health, Education and Welfare.

Hartley, H.O. and Rao, J.N.K. (1978):Estimation of Nonsampling Variance Components in Sample Surveys. In Namboodiri, N.K. (Ed): Survey Sampling and Measurement, pp. 35-43, Academic Press.

Harvig, H. (1973a): Kontrollkodning av dödsbevis. Memo, Statistics Sweden (In Swedish).

Harvig, H. (1973b): Kontrollkodningsexperiment på blanketter för inskrivningsuppgifter till högre studier. Memo, Statistics Sweden (In Swedish).

Jabine, T.B. and Tepping, B.J. (1973): Controlling the Quality of Occupation and Industry Data. Invited paper presented at the International Statistical Institute meeting in Austria, 1973.

Knaus,R. (n.d.): Syntactically Based Classification from Natural Language Responses. Memo, US Bureau of the Census.

Knaus, R. (1978a): Inference by Semantic Pattern Matching in Industry Classification. Memo, US Bureau of the Census.

Knaus, R. (1978b): Automated Industry Coding - an Artificial Intelligence Approach, Memo, US Bureau of the Census.

Knaus, R. (1979): A Similarity Measure on Semantic Network Nodes. Paper presented at the Classification Society Annual Meeting, Gainesville, Florida, 1979.

Knaus, R. (1983): Methods and Problems in Coding Natural Language Survey Data. American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 51-60.

Lakatos, E. (1977a): Automated I & O Coding. Memo, US Bureau of the Census.

Lakatos, E. (1977b):Computerized Coding of Free Verbal Responses. Memo, US Bureau of the Census.

Linebarger, J.S., Jablin, C., and Davie, W.C. (1976): Dependent versus Independent Verification. Memo, US Bureau of the Census.

Lyberg, L. (n.d.): Beroende och oberoende kontroll av kodning. Rapport nr 4, Forskningsprojektet FEL I UNDERSÖKNINGAR, Stockholms universitet, Stockholm (In Swedish).

Lyberg, L. (1969): On the Formation of Coding Teams in the Case of Independent Verification Under Cost Considerations. Forskningsprojektet FEL I UNDERSÖKNINGAR, Stockholms universitet, Stockholm, rapport nr 18.

Lyberg, L. (1981): Control of the Coding Operation in Statistical Investigations - Some Contributions. Ph.D. Thesis, Urval No. 13, Statistics Sweden.

Lyberg, L. (1983): The Development of Procedures for Industry and Occupation Coding at Statistics Sweden. Statistical Review, pp. 139-156.

Lyberg, L., Nordling, P. and Elmdahl, J. (1973): Kodningskvaliteten i lärarregistret. Memo, Statistics Sweden (In Swedish).

Minton, G. (1968): Some Observations Concerning a Model for Classification Control. Memo, US Bureau of the Census.

Minton, G. (1969): Inspection and Correction Error in Data Processing. Journal of the American Statistical Association, vol 64, pp. 1256-1275.

Minton, G. (1970a): Some Decision Rules for Administrative Applications of Quality Control. Journal of Quality Technology, vol. 2, no 2, pp. 86-98.

Minton, G. (1970b): Comments on Quality Control and Research in Data Processing Programs. Paper presented at the American Society for Quality Control, March 12-13, Arlington, Virginia.

Minton, G. (1972): Verification Error in Single Sampling Inspection Plans for Processing Survey Data. Journal of the American Statistical Association, vol 67, no 337, pp. 46-54.

Olofsson, A. (1976): Kvalitetskontroll av näringsgrenskodningen i AKU hösten -74. Memo, Statistics Sweden (In Swedish).

Olofsson, P.O. (1965): PM beträffande variabiliteten i näringsgrens- och yrkesangivelser vid arbetskraftsundersökningar. Memo, Statistics Sweden (In Swedish).

US Bureau of the Census (1965): United States Censuses of Population and Housing, 1960: Quality Control of Preparatory Operations, Microfilming, and Coding. Washington D.C.

US Bureau of the Census (1972): Evaluation and Research Program of the US Censuses of Population and Housing, 1960: Effects of Coders. Series ER 60, no 9, Washington D.C.

Tidigare nummer av Promemorior från P/STM:

NR

1 Bayesianska idéer vid planeringen av sample surveys. Lars Lyberg (1978-11-01)

2 Litteraturförteckning över artiklar om kontingenstabeller. Anders Andersson (1978-11-07)

3 En presentation av Box-Jenkins metod för analys och prognos av tidsserier. Åke Holmén (1979-12-20)

4 Handledning i AID-analys. Anders Norberg (1980-10-22)

5 Utredning angående statistisk analysverksamhet vid SCB: Slutrapport. P/STM, Analysprojektet (1980-10-31)

6 Metoder för evalvering av noggrannheten i SCBs statistik. En översikt. Jörgen Dalén (1981-03-02)

7 Effektiva strategier för estimation av förändringar och nivåer vid föränderlig population. Gösta Forsman och Tomas Garås (1982-11-01)

8 How large must the sample size be? Nominal confidence levels versus actual coverage probabilities in simple random sampling. Jörgen Dalén (1983-02-14)

9 Regression analysis and ratio analysis for domains. A randomization theory approach. Eva Elvers, Carl Erik Särndal, Jan Wretman och Göran Örnberg (1983-06-20)

10 Current survey research at Statistics Sweden. Lars Lyberg, Bengt Swensson och Jan Håkan Wretman (1983-09-01)

11 Utjämningsmetoder vid nivåkorrigering av tidsserier med tillämpning på nationalräkenskapsdata. Lars-Otto Sjöberg (1984-01-11)

12 Regressionsanalys för f d statistikstuderande. Harry Lütjohann (1984-02-01)

13 Estimating Gini and Entropy inequality parameters. Fredrik Nygård och Arne Sandström (1985-01-09)

14 Income inequality measures based on sample surveys. Fredrik Nygård och Arne Sandström (1985-05-20)

15 Granskning och evalvering av surveymodeller, tiden före 1960. Gösta Forsman (1985-05-30)

16 Variance estimators of the Gini coefficient - simple random sampling. Arne Sandström, Jan Wretman och Bertil Waldén (Memo, Februari 1985)

17 Variance estimators of the Gini coefficient - probability sampling. Arne Sandström, Jan Wretman och Bertil Waldén (1985-07-05)

18 Reconciling tables and margins using least-squares. Harry Lütjohann (1985-08-01)

19 Ersättningens och uppgiftslämnarbördans betydelse för kvaliteten i undersökningarna om hushållens utgifter. Håkan L. Lindström (1985-11-29)

20 A general view of estimation for two phases of selection. Carl-Erik Särndal och Bengt Swensson (1985-12-05)

21 On the use of automated coding at Statistics Sweden. Lars Lyberg (1986-01-16)

Kvarvarande exemplar av ovanstående promemorior kan rekvireras från Elseliv Lindfors, P/STM, SCB, 115 81 Stockholm, eller per telefon 08 7834178

★Garnisonstryckeriet★Stockholm★1986★